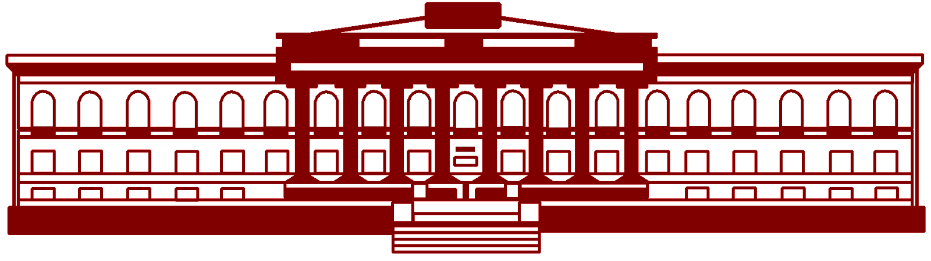
**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**імені ТАРАСА ШЕВЧЕНКА**

****

**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

**Звіт до лабораторної роботи №2**

**з курсу**

**«Data Science та Big Data»**

Студента 4 курсу

Групи ПП-41

Спеціальності 122 «Комп'ютерні науки»

ОП «Прикладне програмування»

Цимбала Дмитра

Викладач:

Білий Р.О.

**Київ – 2022**

## Розвідувальний аналіз даних (EDA). Складання аналітичного звіту

**1. Мета роботи**

Метою лабораторної роботи є отримання практичних навичок виконання розвідувального аналізу даних, використовуючи пакети jupyter, pandas, seaborn. Ознайомлення з методологією складання аналітичного звіту для зовнішнього користувача інформаційного продукту.

**2. Контекст**

### Ви – щойно нанятий data analyst у великій американській компанії, яка працює на ринку нерухомості США. На черговому засіданні ваш бос дав вам завдання зробити аналітичний звіт по цікавому йому сегменту ринку - Нью-Йорку.

### 3. Завдання до виконання

### Виконайте дослідження domain experience стосовно американського ринку нерухомості. Ознайомтесь з декількома прикладами аналітичних продуктів від топових гравців на американському ринку, направлених на інвесторів. Питання, які потрібно опрацювати:

### Як топові компанії на ринку складають звіти по нерухомості?

### Які графіки використовуються для донесення інформації?

### Які співвідношення між якими даними по ринку є показовими для інвесторів / керівників агенцій нерухомості?

### Яка термінологія використовується для опису закономірностей на ринку нерухомості?

### Завантажити файли з даними у папку проекту з посилання:

### <https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page>

### Очистити дані.

### Виконайте розвідувальний аналіз, щоб дізнатися, де є викиди або відсутні значення, вирішіть, як ви їх обробляти, переконайтеся, що дати відформатовані правильно, значення, які ви вважаєте числовими, розглядаються як такі і т.д.

### Виконайте аналіз розвідувальних даних (отриманих результатів) для візуалізації та зіставлення за житловими масивами та за часом. Почніть шукати осмислені закономірності у цьому наборі.

### Зберіть висновки до невеликий звіт для генерального директора (графіки, висновки з текстом у окремому файлі), який потребує належного оформлення висновків, структури тощо.

### Виконання:

# середня ціна по району

average\_price\_by\_borough = all\_data.groupby('BOROUGH')['SALE PRICE'].mean().reset\_index()

plt.bar(average\_price\_by\_borough['BOROUGH'], average\_price\_by\_borough['SALE PRICE'])

plt.gca().get\_yaxis().set\_major\_formatter(ticker.StrMethodFormatter('{x:,.0f}'))

plt.xlabel('Borough')

plt.ylabel('Average selling price')

plt.title('Average selling price by Borough')

plt.show()

### 

### Висновок:

# Райони із найвищою та найнижчою середньою ціною

average\_prices = all\_data.groupby(['BOROUGH', 'NEIGHBORHOOD'])['SALE PRICE'].mean().reset\_index()

highest\_prices = average\_prices.groupby('BOROUGH')['SALE PRICE'].idxmax()

neighborhoods\_with\_highest\_prices = average\_prices.loc[highest\_prices]

lowest\_prices = average\_prices.groupby('BOROUGH')['SALE PRICE'].idxmin()

neighborhoods\_with\_lowest\_prices = average\_prices.loc[lowest\_prices]

neighborhoods\_highest\_labels = [f'{row["NEIGHBORHOOD"]}, {row["BOROUGH"]}' for \_, row in neighborhoods\_with\_highest\_prices.iterrows()]

neighborhoods\_lowest\_labels = [f'{row["NEIGHBORHOOD"]}, {row["BOROUGH"]}' for \_, row in neighborhoods\_with\_lowest\_prices.iterrows()]

# Відобразимо результати на графіку

plt.figure(figsize=(12, 6))

plt.bar(neighborhoods\_highest\_labels, neighborhoods\_with\_highest\_prices['SALE PRICE'], color='red', label='Highest price')

plt.bar(neighborhoods\_lowest\_labels, neighborhoods\_with\_lowest\_prices['SALE PRICE'], color='blue', label='Lowest price')

plt.gca().get\_yaxis().set\_major\_formatter(ticker.StrMethodFormatter('{x:,.0f}'))

plt.xlabel('Borough')

plt.ylabel('Average selling price')

plt.title('Neighborhood with the highest and lowest average price')

plt.xticks(rotation=90)

plt.legend()

plt.show()

### 

### Висновок: як і за минулим графіком, Манхеттен виділився своєю ціною. Але тут видно що минула п’ятірка дорожча за нову яка знаходиться через річку від контенентального Нью-йорку

# Порівняння типів будівель за кількістю продажів

building\_type\_counts = all\_data['BUILDING CLASS CATEGORY'].value\_counts()

threshold = 1500

building\_type\_counts['Other'] = building\_type\_counts[building\_type\_counts < threshold].sum()

building\_type\_counts = building\_type\_counts[building\_type\_counts >= threshold]

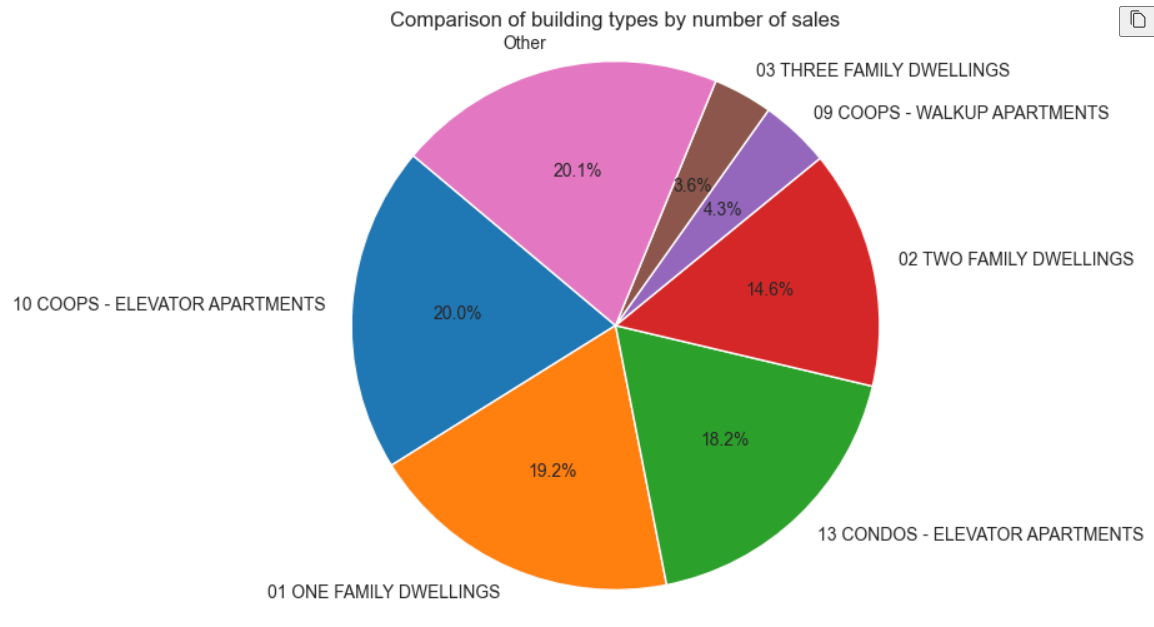
plt.figure(figsize=(6, 6))

plt.pie(building\_type\_counts, labels=building\_type\_counts.index, autopct='%1.1f%%', startangle=140)

plt.axis('equal')

plt.title('Comparison of building types by number of sales')

plt.show()



# Порівняння обсягу продажів за різними податковими класами

sales\_by\_tax\_class = all\_data.groupby('TAX CLASS AT PRESENT')['SALE PRICE'].count().reset\_index()

plt.figure(figsize=(10, 6))

sns.barplot(x='TAX CLASS AT PRESENT', y='SALE PRICE', data=sales\_by\_tax\_class,hue ='TAX CLASS AT PRESENT', palette='plasma')

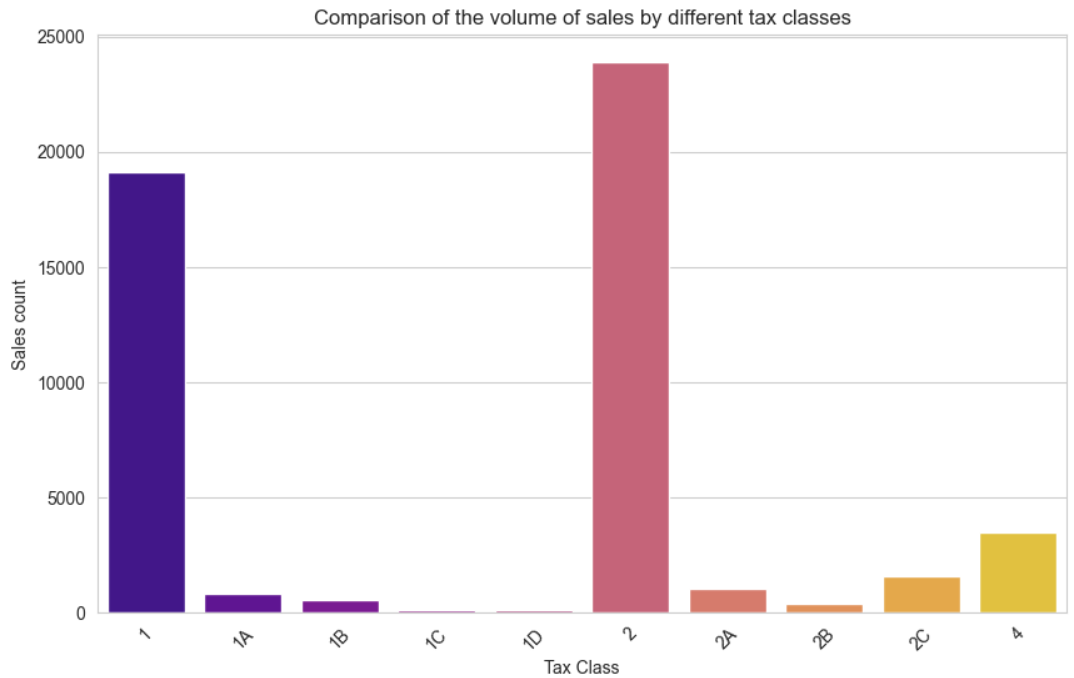
plt.xlabel('Tax Class')

plt.ylabel('Sales count')

plt.title('Comparison of the volume of sales by different tax classes')

plt.xticks(rotation=45)

plt.show()



Висновок: На графіку видно, що у сьогодні найбільшим попитом користуються класи будівель B1, C0, R3, R4, та C1. D4 був надзвичайно популярним раніше, але зараз майже не продається.

# Середня ціна за 1 кв. фут площі у Районі

filtered\_data = all\_data[(all\_data['GROSS SQUARE FEET'] > 0)]

filtered\_data.loc[:,'PRICE PER SQFT'] = filtered\_data['SALE PRICE'] / filtered\_data['GROSS SQUARE FEET']

price\_per\_sqft\_by\_building = filtered\_data.groupby(['BOROUGH'])['PRICE PER SQFT'].mean().reset\_index()

print(price\_per\_sqft\_by\_building)

plt.figure(figsize=(8, 6))

plt.gca().get\_yaxis().set\_major\_formatter(ticker.StrMethodFormatter('{x:,.0f}'))

sns.barplot(data=price\_per\_sqft\_by\_building, x='BOROUGH', y='PRICE PER SQFT', hue='BOROUGH', palette='plasma')

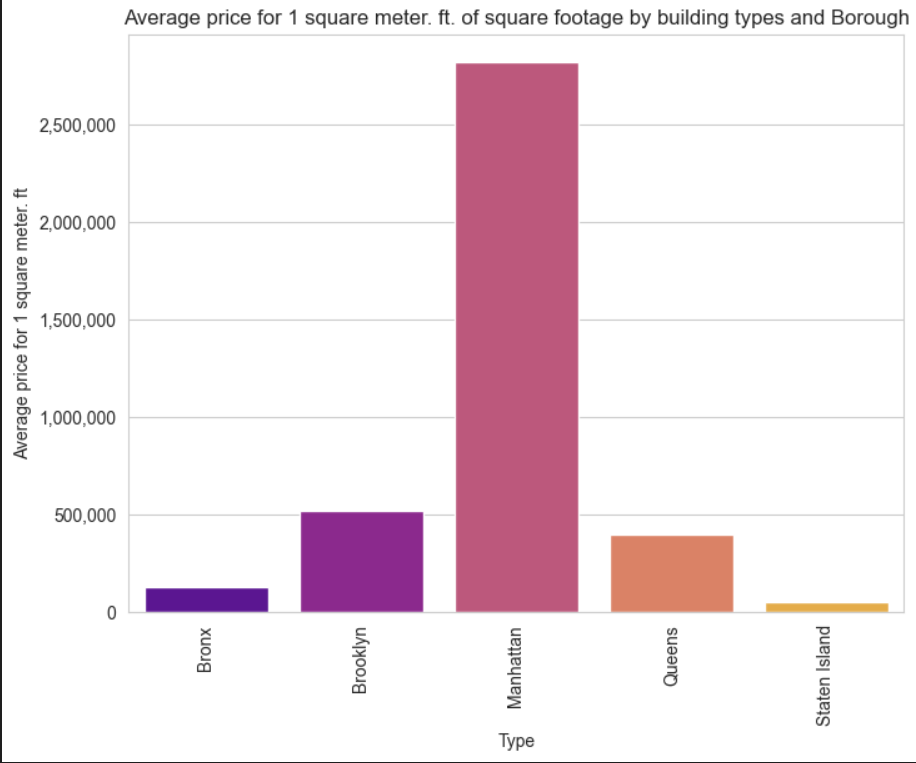
plt.xlabel('Type')

plt.ylabel('Average price for 1 square meter. ft')

plt.title('Average price for 1 square meter. ft. of square footage by building types and Borough')

plt.xticks(rotation=90)

plt.show()



**Висновок:** по старій традиції надорожчий Манхеттен, нейдешевший Staten Island.

# Середня площа будівель за BOROUGH

filtered\_data = all\_data[all\_data['GROSS SQUARE FEET'] > 0]

average\_sqft\_by\_borough = filtered\_data.groupby('BOROUGH')['GROSS SQUARE FEET'].mean().reset\_index()

plt.figure(figsize=(8, 6))

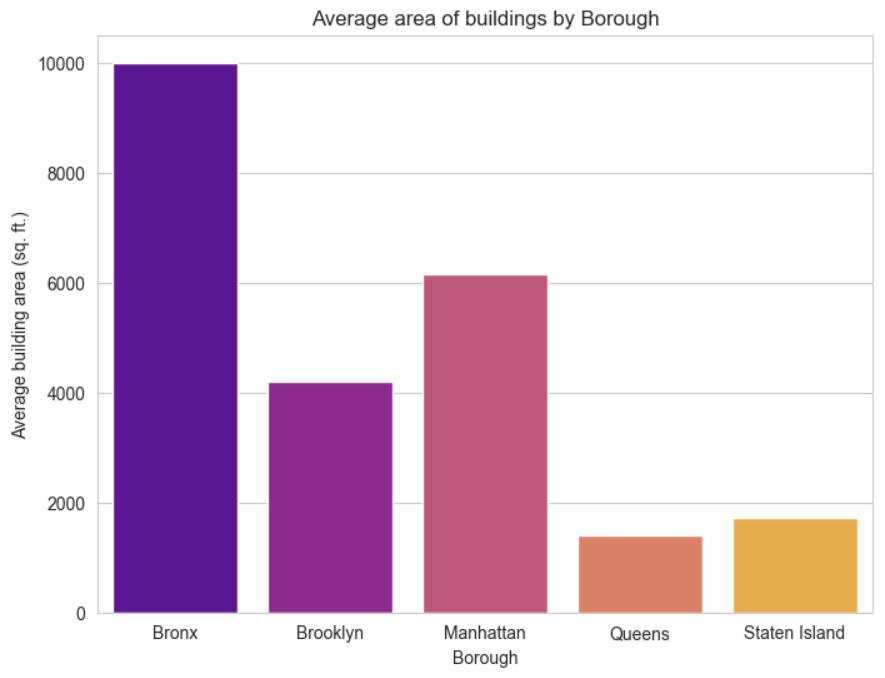
sns.barplot(data=average\_sqft\_by\_borough, x='BOROUGH', y='GROSS SQUARE FEET', hue='BOROUGH', palette='plasma')

plt.xlabel('Borough')

plt.ylabel('Average building area (sq. ft.)')

plt.title('Average area of buildings by Borough')

plt.show()



#Розмір площі, будівель з якою є  найбільа кількість в кожному районі

max\_sqft\_by\_neighborhood = all\_data.groupby('BOROUGH')['GROSS SQUARE FEET'].idxmax()

print(max\_sqft\_by\_neighborhood)

neighborhood\_max\_sqft = all\_data.loc[max\_sqft\_by\_neighborhood, ['BOROUGH', 'GROSS SQUARE FEET']]

print(neighborhood\_max\_sqft)

plt.figure(figsize=(12, 6))

plt.gca().get\_yaxis().set\_major\_formatter(ticker.StrMethodFormatter('{x:,.0f}'))

sns.lineplot(data=neighborhood\_max\_sqft, x='BOROUGH', y='GROSS SQUARE FEET', marker='o')

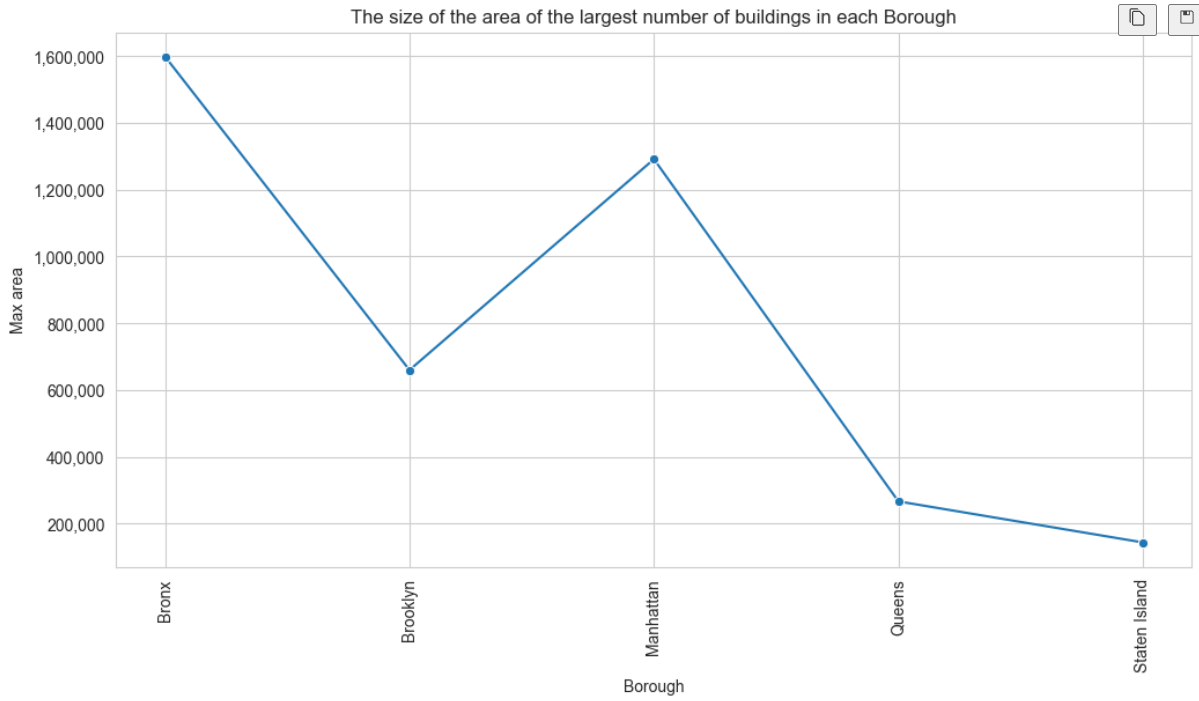
plt.xlabel('Borough')

plt.ylabel('Max area')

plt.title('The size of the area of the largest number of buildings in each Borough')

plt.xticks(rotation=90)

plt.show()



**Висновок:** На графіках бачимо, що у більшості житлових комплексів середня квадратна площа тримається на одному рівні(близько 10000 м^2), але Хантс Поінт має суттєве збільшення(близько 1 580 000 м^2). Можемо зробити висновок, що майже всюди у районі попитом користуються невеликі площі.

**Висновок до всієї роботи:** отримав практичні навички виконання розвідувального аналізу даних, використовуючи пакети jupyter, pandas, seaborn. Ознайомився з методологією складання аналітичного звіту для зовнішнього користувача інформаційного продукту.